Dancing with the sound in edge computing environments

Wangli Hao¹ · Meng Han¹ · Shancang Li² · Fuzhong Li¹

Accepted: 24 September 2021
© The Author(s) 2021

Abstract

Conventional motion predictions have achieved promising performance. However, the length of the predicted motion sequences of most literatures are short, and the rhythm of the generated pose sequence has rarely been explored. To pursue high quality, rhythmic, and long-term pose sequence prediction, this paper explores a novel dancing with the sound task, which is appealing and challenging in computer vision field. To tackle this problem, a novel model is proposed, which takes the sound as an indicator input and outputs the dancing pose sequence. Specifically, our model is based on the variational autoencoder (VAE) framework, which encodes the continuity and rhythm of the sound information into the hidden space to generate a coherent, diverse, rhythmic and long-term pose video. Extensive experiments validated the effectiveness of audio cues in the generation of dancing pose sequences. Concurrently, a novel dataset of audiovisual multimodal sequence generation has been released to promote the development of this field.

Keywords Deep learning · Video sequence generation · Audio guided generation · Multimodal perception · Edge computing

1 Introduction

Audiovisual multimodal generation is a very important research issue in computer vision, and it has a wide range of applications, such as solving the problem of missing modalities, generating zero-shot samples, and inspiring artists to create artistic creations. However, the generation of audiovisual modalities still faces huge challenges due to the large modal discrepancy between visual and audio modalities.

To address the problem of audio-visual multimodal generation, Chen et al. developed a conditional generative adversarial network (GAN) [1] which has several defects. First, the mutual generation of visual-to-audio and audio-to-visual is realized through two independent networks.

Second, the generation of each path is realized through two stages. Specifically, the first stage extracts the discrimination information of the known modality, and the second stage uses the extracted discrimination information to generate the corresponding unknown modalities. In summary, these two phrases limit the efficiency of generation.

Traditional audio-visual multimodal generation has achieved promising results, and most of them are based on a single sample (that is image to sound spectrum or sound spectrum to image). However, as we all know, the visual and audio modalities in the video are both sequential. To further explore the essence of audiovisual multimodal generation and achieve sequence-level generation, we exploit the temporal information in the video for audiovisual multimodal sequence generation. The specific focus is based on the problem of generating dance pose sequence based on sound, referred to as dancing with the sound. Meanwhile, a novel dataset of visual-audio multimodal sequence generation is constructed.

We formalized the task of dancing with the sound into a conditional generation problem (as shown in Fig. 1), and exploring it with various generative models. The key of our task (dancing with the sound) is to use the natural temporal coherence and rhythm of sound signals to generate diverse, reasonable, and smooth dance pose sequences.

✉ Fuzhong Li
hualimengyu@126.com

Shancang Li
Shanshan.Zhao@uwe.ac.uk

¹ School of Software, Shanxi Agricultural University, Jinzhong, Shanxi, China

² Department of Computer Science, University of the West of England, Bristol BS16 1QY, UK

Fig. 1 Process of dance pose sequence generation

The main contributions of this paper can be summarized as follows:

1. We propose the problem of generating dance pose sequence through sound information.
2. We establish a novel visual-audio multimodal sequence generation dataset.
3. We verify the difference between noise and sound information in the process of dance pose sequence generation.
4. We verify the role of sound discrimination and rhythm in the generation of dance pose sequences

The remainder of this paper is organized as follows: Section 2 briefly reviews related works. In Sect. 3, the dance with the sound model is proposed. Experimental results are presented in Sect. 4 and discussions are elaborated in Sect. 5. Section 6 concludes our work.

2 Related works

In this section, we will discuss some research works closely related to our model, including sound to body dynamics, motion transformation and pose to image/video generation respectively.

2.1 Sound to body dynamics

The relationship between voice and facial motion has been extensively studied in some classic works [2, 3]. Recently, the generation of high-quality facial video only from voice information [4] and animation simulation [5, 6] have achieved remarkable results.

Although there is no mature work, exploring only from music to body posture generation. However, there is a number of work related to: multimodal research combining audio and video input, behavioral research on the body's response to music and sound, and research on predicting changes in body posture (such as learning walking and dancing styles from videos) respectively. These methods are summarized below.

2.1.1 Multimodal research combining audio and video input

Multimodal research [7] indicates that combining audio and visual input will generate better results than a single

modality only [8]. For example, facial expression recognition performance with additional voice input will perform better, because the additional information can help identify [9, 10]. Studies by Wang et al. [11] show that the body posture with voice input [12] is estimated with a high accuracy. Multi-speaker tracking [13] combines multimodal information to identify the intention of interaction [14].

2.1.2 Behavioral research on the body's response to music and sound

In dynamic interaction, researchers have studied the relationship between speech and body rhythm. For example, [15, 16] reports that specific audio characteristics are related to the frequency of body movement [17]. Different emotional situations have different types of movement and interaction. For example, regular interactions [17] and interviews [18] have been studied whether there is a correspondence between the pianists' ways of playing specific music, and discovered the way we perceive music-motion correspondence is consistent. Meanwhile, it is shown that this correspondence is flexible, that is, the same music segment may correspond to different music variants.

2.1.3 Research on predicting changes in body posture

Many current researches use long short-term memory (LSTM) to predict and edit future poses from a short motion video or even a photo, such as [19–25]. Nowadays, CNN is usually employed for human pose estimation and recognition [26–30]. These recent works show that it is feasible to learn movement styles from videos [31] and further migrate to animated characters in real time [32]. However, the audio input is not available in their work.

Another complementary research area is predicting audio from video, which is the opposite of predicting video from audio. Some work includes estimating sound from face video for lip reading [33] or utilizing video [34, 35] or photos [36] as input to predict sound objects. Finally, there is work exploring video-driven audio editing [37].

The above mentioned related work does not directly study animations from audio or music to body posture, authors in [38] attempted to verify this possibility. In [38], some researchers have proposed a model that studies the dynamics of sound to body. The model utilizes violin or piano sound as input, outputs the corresponding video pose prediction, and further employs the video pose to make animation. The core idea is to create an animated character based only from audio. This animated character moves like a violinist or pianist when there is music input. They do not expect to reproduce precise body movements. In fact, it is assumed that the transition from speech/music to body

movement is not unique, creating natural body postures are meaningful to music or speech.

Although Ref. [38] explored the possibility from audio to pose and verified their ideas based on experiments, their research focuses on music to violin and piano playing gesture sequences, with simple gestures and small moving space. To solve this problem, we further explored the generation of dance pose sequences from sound.

2.2 Motion transformation

Long-term pedestrian movement can be represented as a transition of a series of motion modes, and the motion mode is a short-term dynamic motion sequence. Yan et al. [39] proposed a motion transform variational autoencoder (VAE) for learning motion sequence generation. Their model jointly learns the feature encoding of one motion mode and the feature transformation that represents the transition from one motion mode to the next. Their model can generate diverse and credible motion sequences based on the same input. Inspired by this model, our work is to learn motion converters for pose sequences.

2.3 Pose to image/video generation

Chan et al. [40] proposed a “follow with me” movement transfer task, that is, transfer their movements to another person according to a given source dancing video. They treated this task as a frame-by-frame image-to-image translation with spatiotemporal smoothing constraints. Specifically, by using the pose representation as an intermediate representation of the source and target, they learned the mapping from the pose image to the appearance of the target subject.

Ma et al. [41] proposed a novel pose-oriented human image generation network (PG2), which can synthesize images of arbitrary pose based on an image of a person and a new pose. PG2 employs specific posture information and consists of two key stages, including posture integration and image refinement. In the first stage, the conditional image and target pose are fed into a UNet-like network to generate an initial but rough image with target pose. In the second stage, a UNet-like generator is trained through adversarial methods to refine the initial blurred image.

Yang et al. [42] extended the pose-oriented image generation network [41], and further proposed a pose-oriented video synthesis model in a decoupled manner. In the first stage, it employs the Pose Sequence Generative Adversarial Network (PSGAN) to generate a pose sequence based on class label through adversarial methods. In the second stage, the Semantic Consistent Generating Adversarial Network (SCGAN) is utilized to generate

video frames from the pose while maintaining the coherent appearance of the input image.

3 Methods

In this section, we will introduce our dance with the sound method in detail.

Due to the appealing motion transition learning capability of motion transformation-VAE (MT-VAE) [39], our model seeks inspiration from it. However, there are differences between them. MT-VAE concerns on the pose prediction under the same modality at different moments. Specifically, the pose at time $t + s$ is predicted from the posture at time t . However, our model focuses on the pose prediction under different modalities at different moments. Concretely, the sound information at time t is utilized the pose at time $t + s$.

Thus, we have made the following improvement to MT-VAE [39] for our work. The input of the model is changed from the pose samples of the same modality at different moments to the pose and sound samples of different modalities at different moments. Consequently, what our model learns is not the motion transition between pose sequences at different moments, but the transition information underlying the pose sequence at the present moment and the sound sample at the next moment.

During the model training phase, our model can generate the next moment pose based on two inputs, the current attitude base and the next moment sound. During the test phase, our model can generate different pose sequences under the guidance of sound at different times. As shown in Fig. 2, our model mainly contains four components, including LSTM encoder, latent encoder, latent decoder, and LSTM decoder respectively.

In Fig. 2, LSTM encoder encodes the pose base at time ($t : t + s$) and music at ($t + s : t + 2s$) into corresponding features named Feature *A* and Feature *B* respectively. Among them, Feature *A* contains the horizontal and vertical coordinates of 18 joint points (the dimension is 36). Feature *B* is the mel-frequency cepstral coefficients (MFCC) feature of the sound. Here, s denotes the step-size of LSTM encoder. The latent encoder encodes the difference between feature *A* and feature *B* into latent space z . The latent decoder generates the pose feature Feature *B*^{*} in the future from latent space z and the current pose feature Feature *A*. The LSTM decoder generates the pose sequence from Feature *B*^{*}.

The loss function of our model during training consists of two terms, including consistency loss function $\mathcal{L}_{Con}$, and motion smoothness function $\mathcal{L}_{Smooth}$ respectively.

$$\mathcal{L}_{Con} = \mathcal{L}_{l2}(P^*, P) \quad (1)$$

Fig. 2 Examples of dataset

where P is the original posture sequence coordinates, and P^* is the generated posture sequence coordinates. This loss function will force the generated coordinates to be as close as possible to the original pose sequence for generating a reasonable pose sequence.

To guarantee the generated pose sequence as smooth as possible, the $\mathcal{L}_{Smooth}$ function is utilized to constrain the second-order difference of the generated pose sequence coordinates and those of the original pose sequence as close as possible, expect to generate a smooth pose sequence.

$$\mathcal{L}_{Smooth} = \frac{1}{K-1} \sum_{t=1}^{K-1} \|m_t^* - m_t\| \quad (2)$$

where K is the total time step, and m_t is the second-order difference term, which can be represented as $m_t = P_{t+1} + P_{t-1} - 2 * P_t$. in which m_t^* has the similar meaning, and it can be denoted as $m_t^* = P_{t+1}^* + P_{t-1}^* - 2 * P_t^*$.

4 Experimental results and analysis

4.1 Setup

To perform our newly proposed task on dance with the sound, a novel dataset including 116 ballet dance recitals is established. For convenience and fair comparison, we synchronize all videos to 24 fps (frame per second). Specifically, 80% of the dataset is randomly chosen as training set and the rest is treated as test set, separately. The samples of the dataset are shown in Fig. 3.

Fig. 3 Examples of dataset

Meanwhile, we employ the openpose [43–46] tool to extract the poses of all video images, as shown in Fig. 4. In Fig. 4, the first row indicates the original image, the second row denotes the original image annotated with 18 joint points, and the third row presents the 18 joint points extracted corresponding to each original image.

Adam is utilized to train the model. The number of LSTM hidden units is 512, the step size is 400 and the batchsize is 16. Furthermore, the learning rate is 0.0001, and the number of iterations is 200 epochs.

4.2 Evaluation of the discrimination of the audio information

Discrimination is essential in the prediction and generation computer vision tasks. If the input music/sound segment is not discriminative, which cannot generate a reasonable dance pose sequence. Consequently, to tackle the task on dancing with music, we verify the discriminative capability of the music from the three aspects, including difference between music-guided and noise-guided pose generations, same pose sequence with different audio sequences and same audio sequence with different pose sequences respectively.

4.2.1 Evaluation of the difference between music-guided and noise-guided pose generations

Most traditional pose generation tasks are based on noise to generate target samples (such as images, etc.). However, these tasks are different from ours. Specifically, most of them focus on the generation of a single sample, and our task concerns the sequence-level generation. Further, the condition of our task is sound rather than noise. To verify

Fig. 4 Examples of pose joint points extracted from the dataset

the effectiveness of our model, we first need to figure out whether the noise can generate a pose sequence. If possible, we further need to explore the difference between the dance pose sequence generated by noise and sound. The result of the comparison of the generated pose sequences from noise and sound are shown in Fig. 5. (a) of Fig. 5 shows the pose sequence generated by noise. (b) of Fig. 5 presents the pose sequence generated by sound. Figure 5 demonstrates that there is almost no diversity in the dance pose generated by noise, indicating that the model of sound conditions has collapsed. On the contrary, the dance pose sequence generated by sound has promising diversity, which shows that music samples have superior discriminativeness than noise. In order to make readers more clearly see the changes in the generated dance pose, we present them with a stepsize 10. All subsequent experimental results are illustrated in this way by default.

4.2.2 Evaluation of the same pose sequence with different audio sequences

To verify the discriminativeness of sound information, we compared the generated dance pose sequences based on the same pose base and sound sample at different moment. The comparison results are shown in Fig. 6. Specifically, the input pose base for all rows of Fig. 6 is sampled from time range $t = 0:200$, the input sound sample for (a), (b), (c) and (d) of Fig. 6 are sampled from time range 0:200, 200:400, 400:600 and 600:800 respectively. From Fig. 6, it is observed that the generated pose sequences based on the same pose base with sound at different moment are different, validating the discriminativeness of the sound information.

Fig. 5 Generated dance pose sequence with noise or music

Fig. 6 Generated dance pose sequence in different time based on the same pose base and different sound. **a** Generated dance pose sequence based on sound at 0:200 **b** Generated dance pose sequence based on sound at 200:400 **c** Generated dance pose sequence based on sound at 400:600 **d** Generated dance pose sequence based on sound at 600:800

4.2.3 Evaluation of the same audio sequence with different pose sequences

To validate the discriminativeness of the sound information from another perspective, we compared the generated dance pose sequences based on the same sound sample and pose base at different moments. The comparison results are shown in Fig. 7. Specifically, the input sound sample for all rows of Fig. 7 is sampled from time range $t=0:200$, the input pose base for (a), (b), (c) and (d) of Fig. 7 are sampled from time range 0:200, 200:400, 400:600 and 600:800 respectively. Figure 7 illustrates that the generated pose sequences based on the same sound sample with different pose base are almost the same, which verifies the discriminativeness of the sound signal from the opposite side.

Fig. 7 Generated dance pose sequence in different time based on the different pose base and the same sound, **a** Generated dance pose sequence based on pose base at 0:200 **b** Generated dance pose sequence based on pose base at 200:400 **c** Generated dance pose sequence based on pose base at 400:600 **d** Generated dance pose sequence based on pose base at 600:800

4.3 Evaluation of the influence of the sound rhythm on the generated dance rhythm

In order to verify the influence of sound rhythm on the generated dance pose rhythm, we compared different models, including M_{slow} , M_{normal} and M_{fast} respectively. Among them, M_{slow} indicates that the model utilizes the audio with a 2 times slower rhythm, which is implemented by performing 2 times upsampling on the original audio sample. M_{normal} denotes the model employs the audio with original rhythm. M_{fast} represents the model uses the audio with a 2 times faster rhythm, which is obtained by executing 2 times downsampling on the original audio sample.

Comparison results are reported in Fig. 8. (a), (b) and (c) of Fig. 8 show the pose sequences generated by M_{slow} , M_{normal} and M_{fast} models respectively. Figure 8 demonstrates that different sound rhythm will generate pose sequences with different rhythms, which further verify the manipulability of the audio information. It should be noted

Fig. 8 Generated dance pose sequence based on music with different rhythms

here that the speed of the rhythm is reflected by the number of modes of pose motion displayed. The M_{slow} model generates fewer pose modes. On the contrary, there are more pose motion modes of M_{fast} model as expected.

4.4 Evaluation of the effective iterative updation

Based on the pose base, our model will generate a reasonable dance pose sequence under the guidance of the sound sample conditions. The step size of the input pose base and sound are both 400, and the generated pose sequence also with the step size of 400. In order to reduce the length of the pose base and without significant performance degradation, we utilize iterative strategy to generate pose sequence. Specifically, assuming that the total step size of the pose base is 400 and the iteration step size is t , the number of iteration generations is $k = (400/t)$ times. That is to say, with the pose step size t , under the guidance of the sound sample with the step size t , a pose sequence of length t is generated. Then based on the generated pose sequence of length t , and the sound samples collected from time $t + 1$ to $2t$, the model will generate the next pose sequence of length t . Iterate successively k times, and concatenating these generated pose sequences together, the model will obtain a dance pose sequence with a total length of 400.

To verify the useful guidance information contained in sound information and the effectiveness of iterative

updation, models with different iterative step sizes, including 1, 4, 10, 50, 80, 200, 400 respectively, are employed respectively. Experimental results are shown in Fig. 9. Among them, (a)–(g) of Fig. 9 present the generated pose sequences with 1, 4, 10, 50, 80, 200, 400 stepsize respectively. illustrates that when the iteration step size is 1, the space coordinates of the generated dance pose sequence are severely offset because of the excessive accumulation of errors. When the iteration step size is 4 and 10, the spatial coordinates of the generated dance pose sequence are more reasonable. When the iteration step is 50, 80, 200, 400 the generated dance pose sequences are much better. Therefore, under the guidance of sound conditions, the iterative update strategy is very effective. When the iterative step length is $1/8$ of the total step length or less, a dance pose sequence almost equivalent to the original model can be obtained, which greatly decreases the training sample size.

4.5 Evaluate the different audio feature representation

To validate the effectiveness of different audio feature representations, we design two groups of experiments. One group is to assess the smoothness and other is to verify the discriminative capability of different audio features.

Concerning the first group, we compare models with various audio features, including mfcc, logmfcc, mel, logmel, to evaluate the smoothness of different audio feature representations. Figure 10 exhibits the experimental results. We find that, on the one hand, features including mel are smoother compared to their corresponding mfcc variants ($\text{mel} > \text{mfcc}$; $\text{logmel} > \text{logmfcc}$). In addition, features without log transformation are smoother ($\text{mel} > \text{logmel}$, $\text{mfcc} > \text{logmfcc}$).

For the second group, we compare models with various audio features, including mfcc, logmfcc, mel, logmel, to assess the discriminative capability of different audio feature representations. Experimental results are displayed in Fig. 10. The mel feature has the worst discriminative, which is because it is infeasible to perform iterative pose prediction when the iterative step is smaller than 200.

Above all, we figure out that the smoothness and the discriminative of audio features are contrary and we chose the trade-off features, mfcc, as our default audio feature representation.

5 Discussions

Most traditional pose or motion prediction are formed as a conditional generation problem. However, the conditions utilized are usually noise, motion label and previous

Fig. 9 Generated dance pose sequence based on different iteration stepsize

Fig. 10 Generated dance pose sequence with different audio features, **a** Logmel **b** Logmfcc **c** Mel **d** MFCC

pose/motion sequence, which generated pose/motion sequence with short length. To generate long term pose sequence, we leverage the sound as condition. We believe an assumption that the sound is sufficiently discriminative to facilitate the model to generate long-term pose sequences. Several experiments have verified our hypothesis, including the generated pose sequences based on the same sound sample with different pose base are almost the same, the generated pose sequences based on the same pose base and different sound sequences are different, and the dance pose sequence generated by sound has much more diversity than that of noise respectively.

Moreover, to further mining the discriminativeness of sound information, we reduce the length of pose base and utilize iterative strategy to generate pose sequence. Experimental results demonstrate that when the length of pose base is 1/8 of the total step length or less, the generated dance pose sequence almost equivalent to that of the original model, which greatly decreases the training sample size. This result further illustrates the discriminativeness of sound information from another aspect.

Additionally, the sound signal has several promising properties, such as internal coherence, rhythm, etc., which endow the generated pose sequence with appealing coherence and rhythmic. The above experiment in Sect. 4.3 verifies the effectiveness of these sound properties in the pose sequence generation.

6 Conclusions

This paper proposes a novel dancing with the sound task, aims at exploring the feasibility of predicting long-term pose sequences with the aid of sound information and the possibility of cross-modal generation. Further, a model based on VAE framework has been proposed to tackle this task, which leverages the sound as input condition and output the corresponding dancing pose sequence. Additionally, a new dancing with the sound dataset has been established. Numerous ablation studies have verify the effectiveness of our model and the dance with the sound task. Specifically, the discriminativeness of sound information enables the model to generate a long-term pose sequence, and the natural temporal coherence and rhythm of sound information allow the generated pose sequence with promising plausibility and the rhythmic.

Funding This work was supported by Shanxi Province Higher Education Innovation Project of China 2020L0154, Intelligent Information Processing Shanxi Provincial Key Laboratory Open Project Fund (CICIP2021005), Shanxi Agricultural University Academic Recovery Research Project (2020xshf38), Shanxi Key Research and Development Program (201703D221033-3), and NSFC of China under Grants 61702317.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Chen, L., Srivastava, S., Duan, Z., & Xu, C. (2017). Deep cross-modal audio-visual generation. In *Proceedings of the on thematic workshops of ACM multimedia 2017* (pp. 349–357). ACM.
2. Brand, M. (1999). Voice puppetry. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (pp. 21–28). ACM Press/Addison-Wesley Publishing Co.
3. Bregler, C., Covell, M., & Slaney, M. (1997). Video rewrite: Driving visual speech with audio. *Siggraph*, 97, 353–360.
4. Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4), 95.
5. Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., et al. (2017). A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4), 93.
6. Karras, T., Aila, T., Laine, S., Herva, A., & Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4), 94.
7. Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1–2), 116–134.
8. Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., & Rigoll, G. (2013). LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2), 153–163.
9. Chen, L.S.-H., & Huang, T. S. (2000). *Joint processing of audio-visual information for the recognition of emotional expressions in human–computer interaction*. Citeseer.
10. Sebe, N., Cohen, I., & Huang, T. S. (2005). Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision* (pp. 387–409). World Scientific.
11. Wang, S. B., & Demirdjian, D. (2005). Inferring body pose using speech content. In *Proceedings of the 7th international conference on multimodal interfaces* (pp. 53–60). ACM.
12. Ouyang, W., Chu, X., & Wang, X. (2014). Multi-source deep learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2329–2336).
13. Ban, Y., Girin, L., Alameda-Pineda, X., & Horaud, R. (2017). Exploiting the complementarity of audio and visual data in multi-speaker tracking. In *Proceedings of the IEEE international conference on computer vision* (pp. 446–454).
14. Schwarz, J., Marais, C. C., Leyvand, T., Hudson, S. E., & Mankoff, J. (2014). Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3443–3452). ACM.
15. Dittmann, A. T. (1972). The body movement-speech rhythm relationship as a cue to speech encoding. In *Studies in dyadic communication* (pp. 135–152).
16. Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview.
17. Boomer, D. S., & Dittman, A. P. (1964). Speech rate, filled pause, and body movement in interviews. *Journal of Nervous and Mental Disease*, 139, 324–327.
18. Dittmann, A. T., & Llewellyn, L. G. (1969). Body movement and speech rhythm in social conversation. *Journal of Personality and Social Psychology*, 11(2), 98–106.
19. Fragkiadaki, K., Levine, S., Felsen, P., & Malik, J. (2015). Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision* (pp. 4346–4354).

20. Martinez, J., Black, M. J., & Romero, J. (2017). On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2891–2900).
21. Habibie, I., Holden, D., Schwarz, J., Yearsley, J., & Komura, T. (2017). A recurrent variational autoencoder for human motion synthesis. In *BMVC17*.
22. Holden, D., Saito, J., & Komura, T. (2016). A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4), 138.
23. Li, Z., Zhou, Y., Xiao, S., He, C., & Li, H. (2017). Auto-conditioned LSTM network for extended complex human motion synthesis (vol. 3). arXiv preprint [arXiv:1707.05363](https://arxiv.org/abs/1707.05363)
24. Walker, J., Marino, K., Gupta, A., & Hebert, M. (2017). The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision* (pp. 3332–3341).
25. Chao, Y.-W., Yang, J., Price, B., Cohen, S., & Deng, J. (2017). Forecasting human dynamics from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 548–556).
26. Insafutdinov, E., & Schiele, B. (2016). Dense-cnn: Fully convolutional neural networks for human body pose estimation. Ph.D. dissertation, Universität des Saarlandes Saarbrücken.
27. Belagiannis, V., & Zisserman, A. (2017). Recurrent human pose estimation. In *2017 12th IEEE international conference on automatic face and gesture recognition (FG 2017)* (pp. 468–475). IEEE.
28. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., & Schiele, B. (2016). Deeppercut: A deeper, stronger, and faster multi-person pose estimation model. In *European conference on computer vision* (pp. 34–50). Springer.
29. Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4724–4732).
30. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).
31. Liu, C. K., Hertzmann, A., & Popović, Z. (2005). Learning physics-based motion style with nonlinear inverse optimization. *ACM Transactions on Graphics (TOG)*, 24(3), 1071–1081.
32. Grochow, K., Martin, S. L., Hertzmann, A., & Popović, Z. (2004). Style-based inverse kinematics. *ACM Transactions on Graphics (TOG)*, 23(3), 522–531.
33. Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3444–3453). IEEE.
34. Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G. J., Durand, F., & Freeman, W. T. (2014). The visual microphone: Passive recovery of sound from video.
35. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., & Freeman, W. T. (2016). Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2405–2413).
36. Soler, M., Bazin, J.-C., Wang, O., Krause, A., & Sorkine-Hornung, A. (2016). Suggesting sounds for images from video collections. In *European conference on computer vision* (pp. 900–917). Springer.
37. Liao, Z., Yu, Y., Gong, B., & Cheng, L. (2015). Audeosynth: Music-driven video montage. *ACM Transactions on Graphics (TOG)*, 34(4), 68.
38. Shlizerman, E., Dery, L., Schoen, H., & Kemelmacher-Shlizerman, I. (2018). Audio to body dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7574–7583).
39. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., & Lee, H. (2018). Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 265–281).
40. Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2018). Everybody dance now. arXiv preprint [arXiv:1808.07371](https://arxiv.org/abs/1808.07371)
41. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Gool, L. (2017). Pose guided person image generation. In *Advances in neural information processing systems* (pp. 406–416).
42. Yang, C., Wang, Z., Zhu, X., Huang, C., Shi, J., & Lin, D. (2018). Pose guided human video generation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 201–216).
43. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2018). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. arXiv preprint [arXiv:1812.08008](https://arxiv.org/abs/1812.08008)
44. Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.
45. Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.
46. Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *CVPR*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Wangli Hao is currently a lecturer at Shanxi Agricultural University and director of the Academic Affairs and Scientific Research Office. Her current research interests include artificial intelligence, multi-modal perception. She received her bachelor's degree in Shanxi Agricultural University in 2011. After that, She received her master's degree in Beijing Institute of Technology in 2014. Finally, she received her Ph.D.'s degree in Institute of Automation, Chinese Academy of Sciences.

Meng Han is currently a lecturer at Shanxi Agricultural University. His current research interests include artificial intelligence, multi-modal perception. He received his bachelor's degree in Shanxi Agricultural University in 2011. After that, he received her master's degree in Shanxi Agricultural University in 2017.

Dr. Shancang Li (B.Eng., M.Eng., Ph.D.) is a senior lecturer in Network Forensics at UWE. His current research interests include network forensics, device security, wireless sensor networks, Internet of Things, and lightweight cryptography over IoT.

Fuzhong Li is currently the dean of the School of Software of Shanxi Agricultural University, Ph.D., professor, and doctoral supervisor. The only young academic leader in agricultural economics in Shanxi Province, a visiting scholar at Oberlin University and Michigan State University in the United States (2006–2007), a member of a delegation of one hundred young Chinese scientific and technological workers to France.